QFlow 2.0: Quantum dot data for machine learning

Justyna P. Zwolak,[1,*] Joshua Ziegler,[1] Sandesh S. Kalantre,[2,3] and Jacob M. Taylor[1,2,3]

[1] *National Institute of Standards and Technology, Gaithersburg, MD 20899, USA*
[2] *Joint Quantum Institute, University of Maryland, College Park, MD 20742, USA*
[3] *Joint Center for Quantum Information and Computer Science, University of Maryland, College Park, MD 20742, USA*

Arrays of quantum dots (QDs) are a promising candidate system to realize scalable, coupled qubits systems and serve as a fundamental building block for quantum computers. However, establishing a stable configuration of electrons in space is a non-trivial task achieved via electrostatic confinement, band-gap engineering, and dynamically adjusted voltages on nearby electrical gates. A key task is to determine a good set of control parameters (gate voltages) to achieve a desired charge configuration—in both number and location—for a successful experiment.

In recent years, many research groups working with QDs began to implement machine learning (ML) techniques to automate the QD tuning task. This dataset, consisting of both simulated and experimental QD measurement data, was established to enable development and benchmarking of ML tools for automation of QD experiments.

CONTENTS

REFERENCES

If you use the QFlow 2.0 dataset in your research, please cite:

> J. P. Zwolak, J. Ziegler, S. S. Kalantre, and J. M. Taylor (2022), QFlow 2.0: Quantum dot data for machine learning, National Institute of Standards and Technology, https://doi.org/10.18434/mds2-1894 (Accessed 20YY-MM-DD).

as well as

> J. Ziegler, T. McJunkin, E. S. Joseph, S. S. Kalantre,0 B. Harpt, D. E. Savage, M. G. Lagally, M. A. Eriksson, J. M. Taylor, and J. P. Zwolak. *Toward Robust Autotuning of Noisy Quantum Dot Devices.* Phys. Rev. Applied **17** (2), 024069 (2022).

A complete list of references related to the QFlow dataset:

[1] S. S. Kalantre, J. P. Zwolak, S. Ragole, X. Wu, N. M. Zimmerman, M. D. Stewart, and J. M. Taylor. *Machine Learning techniques for state recognition and auto-tuning in quantum dots.* npj Quantum Information **5** (6): 1–10 (2019). doi:10.1038/s41534-018-0118-7

[2] J. P. Zwolak, S. S. Kalantre, X. Wu, S. Ragole, and J. M. Taylor. *QFlow lite dataset: A machine-learning approach to the charge states in quantum dot experiments.* PLoS ONE **13** (10): e0205844 (2018). doi: 10.1371/journal.pone.0205844

* jpzwolak@nist.gov

[3] J. P. Zwolak, T. McJunkin, S. S. Kalantre, J. P. Dodson, E. R. MacQuarrie, D. E. Savage, M. G. Lagally, S. N. Coppersmith, M. A. Eriksson, and J. M. Taylor. *Autotuning of double-dot devices in situ with machine learning.* Phys. Rev. Applied **13** (3): 034075 (2020). doi:10.1103/PhysRevApplied.13.034075

[4] J. P. Zwolak, S. S. Kalantre, T. McJunkin, B. J. Weber, and J. M. Taylor. *Ray-based classification framework for high-dimensional data.* Proceedings of Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada [December 11, 2020] (2020). doi:10.48550/arXiv.2010.00500

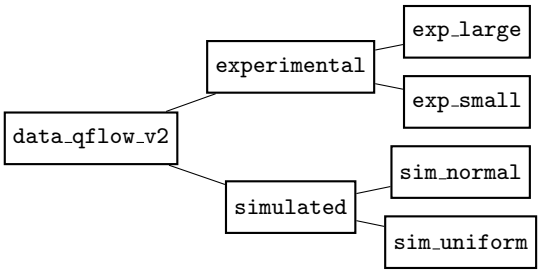
[5] J. P. Zwolak, T. McJunkin, S. S. Kalantre, S. F. Neyens, E. R. MacQuarrie, M. A. Eriksson, and J. M. Taylor. *Ray-based framework for state identification in quantum dot devices.* PRX Quantum **2** (2): 020335 (2021). doi:10.1103/PRXQuantum.2.020335

[6] J. Darulová, M. Troyer, and M. C. Cassidy. *Evaluation of synthetic and experimental training data in supervised machine learning applied to charge-state detection of quantum dots.* Mach. Learn.: Sci. Technol. **2** (4): 045023 (2021). doi:10.1088/2632-2153/ac104c

[7] J. Ziegler, T. McJunkin, E. S. Joseph, S. S. Kalantre, B. Harpt, D. E. Savage, M. G. Lagally, M. A. Eriksson, J. M. Taylor, and J. P. Zwolak. *Toward Robust Autotuning of Noisy Quantum Dot Devices.* Phys. Rev. Applied **17** (2): 024069 (2022). doi:10.1103/PhysRevApplied.17.024069

QFLOW 2.0: DATASET STRUCTURE

The QFlow 2.0 dataset contains the following files:



Experimental data structure

There are two types of experimental files stored in separate folders:

`exp_small`**:** A dataset of 756 small 2D scans, ranging from 30×30 pixels to 60×60 pixels with 1 mV- to 2 mV-per-pixel resolution.

`exp_large`**:** A dataset of 12 large 2D scans, ranging from 126×126 pixels to 401×401 pixels with 1 mV- to 2 mV-per-pixel resolution.

Information about each experimental measurement is stored as an individual NumPy file. Each NumPy file in the `experimental_data` dataset contains the charge sensor data (`'sensor'`), the voltage range for each axis over which the measurement was performed (`'x'` and `'y'`, with `'x'` being the dominant measurement direction), as well as information about the device used in measurement and the cooldown (`'dataset'`). The data was acquired over two cooldowns of one of the devices (indicated as 0 and 1) and a single cooldown of a second device (indicated as 2). The units are not included in the data but they are provided here for completeness. In addition, the `exp_small` includes the human-assigned labels.

Each data file is stored as a dictionary with the following elements (keys) with the data type of each element in the dictionary given in the brackets:

`'x'`: voltage range for the first gate (in volts) [1D NumPy array],

`'y'`: voltage range for the second gate (in volts) [1D NumPy array],

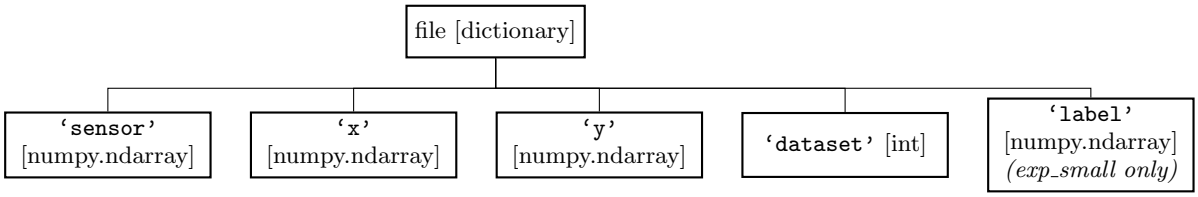
`'sensor'`: the charge sensor data (in amperes) [2D NumPy array],

'dataset': indication of the device and cooldown (when appropriate) (0, 1, or 2) [integer],

'label': the state label for the data [1D NumPy array] (only in data in the exp_small folder).

To see the full data structure of each file see Fig. 1.

FIG. 1. The generic data structure tree for the files. The data type is given in brackets.



Simulated data structure

There are two types of simulated noisy data stored in separate folders (each containing 10 HDF5 files):

sim_normal: A dataset of 1.15×10^5 simulated noisy measurements generated by fixing the relative magnitudes of white noise, $1/f$ noise, and sensor jumps and varying the magnitudes together in a normal distribution. The means of the magnitudes are set to 1.5 times the optimized values and the standard deviation is one-third of each magnitude's value (0.5 of the optimized value). Each file includes 1.15×10^4 simulated measurements, with the noise varied in the same distribution. The file naming convention in this folder is normal_1.5m_0.5std_noisy_data_*X*.hdf5, where $X = 0, \ldots, 9$ indicates the corresponding subset of the data.

sim_uniform: A dataset of 1.15×10^5 simulated noisy measurements with varying amounts of noise added. The magnitudes of all noises that negatively affect the signal-to-noise ratio (sensor jumps, $1/f$, and white noise) are varied together uniformly from 0 to 7 times the optimized noise magnitudes while the dot jumps noise variation is kept within the 1 % used to establish the sim_normal dataset. Each file includes 1.15×10^4 simulated with noise level sampled uniformly over a range of noise levels, with consecutive files being increasingly noisy at increments of 0.7 of the optimized noise composition. The file naming convention in this folder is uniform_noisy_data_quantile_*X*.hdf5, where X indicates the range of noises used in a given subset and takes on the form m-(m+0.70) for $m = 0.00, \ldots, 6.30$.

There are three additional files in the experimental folder. The bn_noCP_0-7uniform_info.csv and bn_noCP_1.5m_0.5std_info.csv files contain information about the noisy simulated data (three columns per each file in the in a particular file in the sim_normal and sim_uniform folders): the noise magnitude used to augment the data and the x and y size after augmentation (note that implementing certain types of noise necessitates slight changes to the file size). The source file information is also included. The noiseless_data.hdf5 file contains physical parameters used to simulate the noiseless measurements as well as the simulated charge sensor and state maps.

Each entry in the hdf5 files in the simulated folder is stored as a dictionary-like object called a group with the following elements (keys) in it (the type of each element is given in the brackets):

'V_P1_vec': voltage range for the first plunger [1D HDF5 dataset],

'V_P2_vec': voltage range for the second plunger [1D HDF5 dataset],

'output': the simulated data [HDF5 group]

In addition, the noiseless_data.hdf5 file contains information about the physical parameters used to simulate each noiseless measurement:

'physics': physical parameters of the device [HDF5 group]

and the files in the sim_normal and sim_uniform folders contain information about the level of the noise added:

'noise_level': the level of the simulated noise [float].

The full data structure of each noiseless file entry in noiseless_data.hdf5 is shown in Fig. 2. The noisy files entries are shown in Fig. 3. For details about the 'output' HDF5 group refer to Table I, and for details about the 'physics' HDF5 group see Table II.

FIG. 2. The generic data structure tree for the entries in the noiseless data HDF5 file. The data type is given in brackets. The simulation outcome data is highlighted in gray. For details about the 'output' data HDF5 group see Table I. For details about the 'physics' data HDF5 group see Table II.

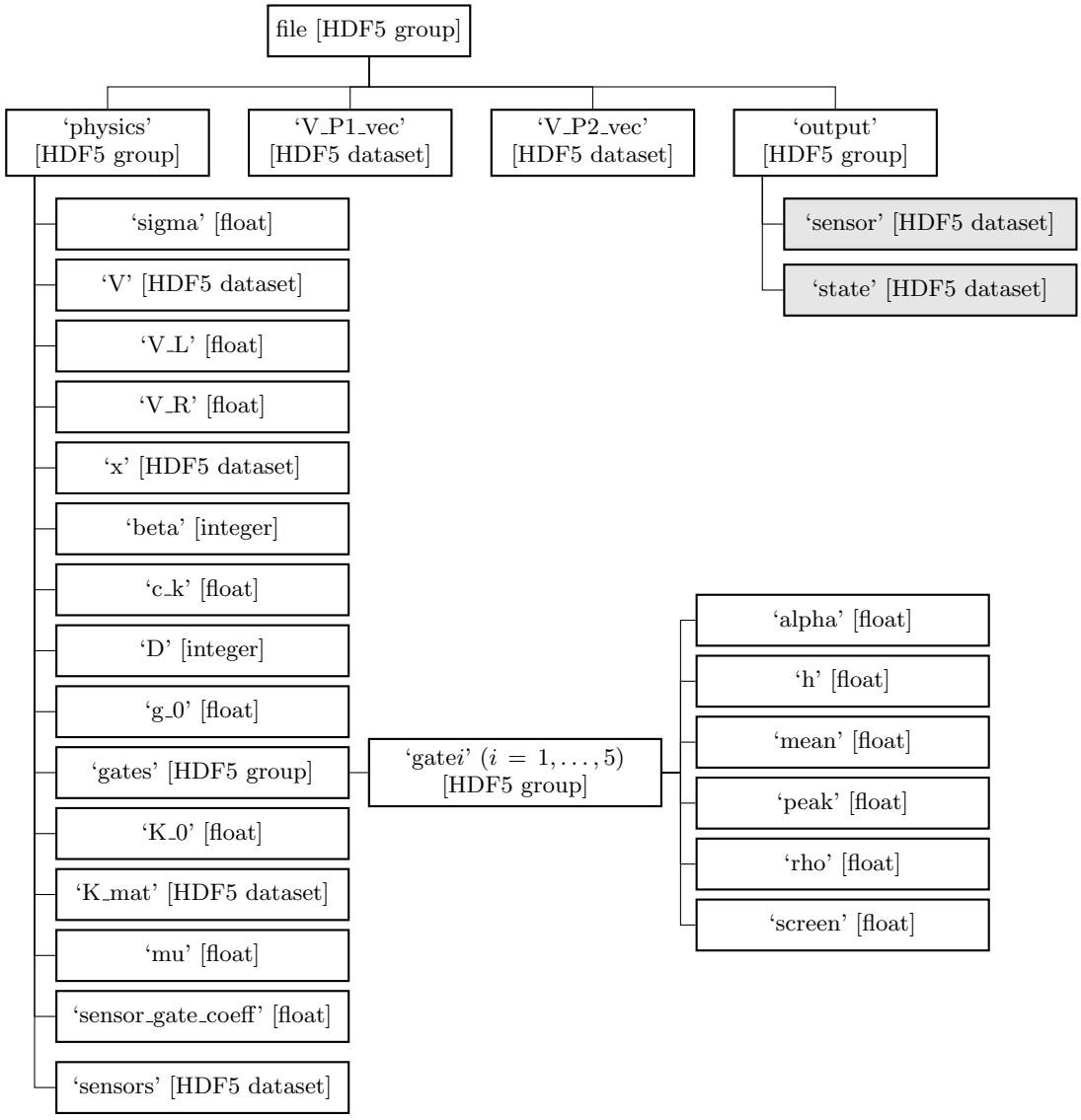


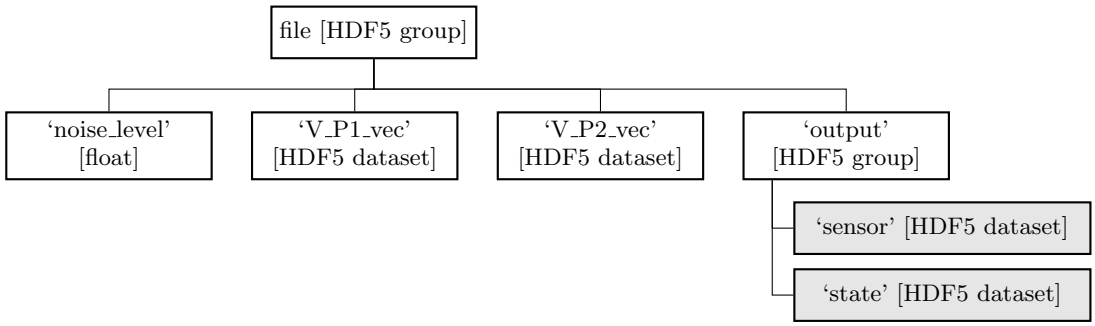
TABLE I. 'output' is a HDF5 group of 2D HDF5 datasets storing the simulated data for each point in the plunger voltage space, as defined by vectors 'V_P1_vec' and 'V_P2_vec'.

Key	Description	Type
'sensor'	the output of the charge sensor evaluated as the Coulomb potential at the sensor location (with artificial noise added if in the noisy sensor data)	float
'state'	the label determining the state of the device, distinguishing between no dot (0), left dot (0.5), central dot (1), right dot (1.5), and a double dot (2)	float

QFLOW LITE: DATASET STRUCTURE

Each NumPy file contains information about a single simulated device, such as physical parameters used in the simulation and the output of the simulation. The units are not included in the data but they are provided here for completeness. Each file is stored as a dictionary with the following five elements (keys) in it (the type of each element in the dictionary is given in the brackets):

FIG. 3. The generic data structure tree for the entries in the noisy data HDF5 files. The data type is given in brackets. The simulation outcome data is highlighted in gray. For details about the 'output' data HDF5 group see Table III.



'type': 'V_P_map' – information about what data is in the file [string],

'V_P1_vec': voltage range for the first plunger (0 to 0.4 V) [(100,) numpy.array],

'V_P2_vec': voltage range for the second plunger (0 to 0.4 V) [(100,) numpy.array],

'output': the simulated data [list];

'physics': physical parameters of the device [dictionary];

To see the full data structure of each file see Fig. 4. For detail about the 'output' refer to Table III and for details about the 'physics' see Table IV.

TABLE II. 'physics' is a HDF5 group with physical parameters of the device. Fixed values are given explicitly. Varied parameters were randomly sampled from a Gaussian distribution with the given mean value μ and standard deviation set to $0.05\,\mu$ (unless stated otherwise).

Key	Description	Value
sigma	softening parameter	3.0 nm
V	potential profile	V(x)
V_L	voltage applied to left lead	50 µV
V_R	voltage applied to right lead	−50 µV
x	linear array spanning the size of the device	$(-60, 60)$ nm
beta	effective temperature used for self-consistent calculation of the electron density $n(x)$	$1000\,(\mathrm{eV})^{-1}$
c_k	kinetic term for the 2DEG	$\langle 1\,\mathrm{meV\,nm}\rangle$
D	dimension of the problem to be used in the electron density integral, (only when polylogarithm function is used to calculate the electron density, for a 2DEG a direct analytic integral of the Fermi function is used)	2
g_0	coefficient of the density of states	$\langle 1.0\,(\mathrm{eV\,nm})^{-1}\rangle$
gates	the HDF5 group of parameters defining each of the five gates:	
	alpha: lever arm (same for all gates)	$\langle 1.0\rangle$
	h: distance of the gate from the electron density (same for all gates)	$\langle 50.0\,\mathrm{nm}\rangle$
	mean: position of the gate along linear array	
	- for gate 1	$\langle -40\,\mathrm{nm}\rangle$
	- for gate 2	$\langle -20\,\mathrm{nm}\rangle$
	- for gate 3	$\langle 0\,\mathrm{nm}\rangle$
	- for gate 4	$\langle 20\,\mathrm{nm}\rangle$
	- for gate 5	$\langle 40\,\mathrm{nm}\rangle$
	peak: potential at the location of the electrons	
	- for gates 1, 3 and 5	$\langle 0.2\,\mathrm{mV}\rangle$
	- for gates 2 and 4	$\langle -0.4\,\mathrm{mV}\rangle$
	rho: radius of the cylindrical gate (same for all gates)	$\langle 5.0\,\mathrm{nm}\rangle$
	screen: the screening length (same for all gates)	$\langle 20.0\,\mathrm{nm}\rangle$
K_0	the strength of the Coulomb interaction	$\langle 10\,\mathrm{meV}\rangle$
K_mat	the Coulomb interaction matrix	K_mat(x,K_0,sigma)
mu	Fermi level (assumed to be equal for both leads)	0.1eV
sensor_gate_coeff	weight applied while including the potential of the gate in calculating the sensor output	0.1
sensors	the position of the charge sensor in the 2DEG plane, stored as (horizontal position with respect to the center of the device, vertical position with respect to the dots which are assumed to be located on the x-axis)	$[(-20, 50)]$ nm

TABLE III. 'output' is a list of dictionaries storing the simulated data for each point in the plunger voltage space, as defined by vectors 'V_P1_vec' and 'V_P2_vec'. There is 10 000 data points (dictionaries), each with four variables defined in the table.

Key	Description	Type
'charge'	the information about the number of charges on each dot (with a default value 0 for short circuit and a barrier)	tuple
'current'	current through the device at infinitesimal bias	float
'sensor'	the output of the charge sensor evaluated as the Coulomb potential at the sensor location	list
'state'	the label determining the state of the device, distinguishing between a single dot (1), a double dot (2) , a short circuit (-1) and a barrier (0)	integer

FIG. 4. The generic data structure tree for the files. The data type is given in brackets. The simulation outcome data is highlighted in gray. For a data dictionary see Table III and Table IV.

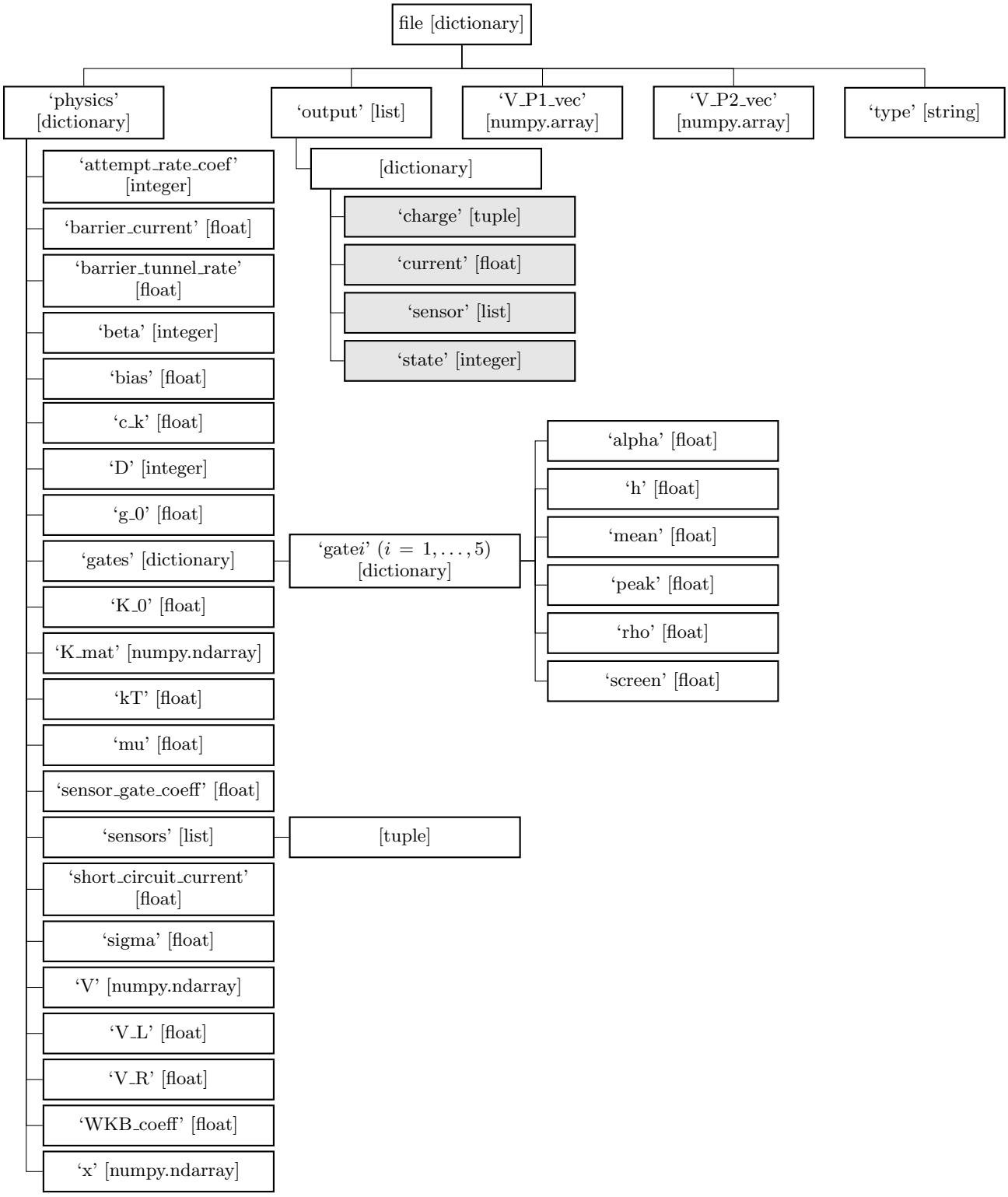


TABLE IV. 'physics' is a dictionary with physical parameters of the device. Fixed values are given explicitly. Varied parameters were randomly sampled from a Gaussian distribution with the given mean value μ and standard deviation set to $0.05\,\mu$ (unless stated otherwise).

Key	Description	Value
attempt_rate_coef	controls the strength of the attempt rate factor	1
barrier_current	a scale for the current set to the device when in barrier mode	1 arb. unit
barrier_tunnel_rate	a tunnel rate set when the device is in barrier mode while calculating the tunnel probability	10.0
beta	effective temperature used for self-consistent calculation of the electron density $n(x)$	$1000\,(\mathrm{eV})^{-1}$
bias	difference in the chemical potential between source and drain	100 eV
c_k	kinetic term for the 2DEG	$\langle 1\,\mathrm{meV\,nm}\rangle$
D	dimension of the problem to be used in the electron density integral, (only when polylogarithm function is used to calculate the electron density, for a 2DEG a direct analytic integral of the Fermi function is used)	2
g_0	coefficient of the density of states	$\langle 1.0\,(\mathrm{eV\,nm})^{-1}\rangle$
gates	the dictionary of parameters defining each of the five gates:	
	alpha: lever arm (same for all gates)	$\langle 1.0\rangle$
	h: distance of the gate from the electron density (same for all gates)	$\langle 50.0\,\mathrm{nm}\rangle$
	mean: position of the gate along linear array	
	- for gate 1	$\langle -40\,\mathrm{nm}\rangle$
	- for gate 2	$\langle -20\,\mathrm{nm}\rangle$
	- for gate 3	$\langle 0\,\mathrm{nm}\rangle$
	- for gate 4	$\langle 20\,\mathrm{nm}\rangle$
	- for gate 5	$\langle 40\,\mathrm{nm}\rangle$
	peak: potential at the location of the electrons	
	- for gates 1, 3 and 5	$\langle 0.2\,\mathrm{mV}\rangle$
	- for gates 2 and 4	$\langle -0.4\,\mathrm{mV}\rangle$
	rho: radius of the cylindrical gate (same for all gates)	$\langle 5.0\,\mathrm{nm}\rangle$
	screen: the screening length (same for all gates)	$\langle 20.0\,\mathrm{nm}\rangle$
K_0	the strength of the Coulomb interaction	$\langle 10\,\mathrm{meV}\rangle$
K_mat	the Coulomb interaction matrix	K_mat(x,K_0,sigma)
kT	temperature of the system used in the transport calculations	50 µK
mu	Fermi level (assumed to be equal for both leads)	0.1eV
sensor_gate_coeff	weight applied while including the potential of the gate in calculating the sensor output	0.1
sensors	the position of the two charge sensors in the 2DEG plane, stored as (horizontal position with respect to the center of the device, vertical position with respect to the dots which are assumed to be located on the x-axis)	$[(-20, 50), (20, 50)]$ nm
short_circuit_current	an arbitrary high current value given to the device when in short circuit mode	100 arb. unit
sigma	softening parameter	3.0 nm
V	potential profile	V(x)
V_L	voltage applied to left lead	50 µV
V_R	voltage applied to right lead	−50 µV
WKB_coeff	the strength of WKB tunneling	0.5
x	linear array spanning the size of the device	$(-60, 60)$ nm